TEXT2EQ: HUMAN-IN-THE-LOOP CO-CREATION INTERFACE FOR EQ

Annie Chu **Hugo Flores García** **Patrick O'Reilly** **Bryan Pardo**
Department of Computer Science, Northwestern University, USA
anniechu@u.northwestern.edu

ABSTRACT

We introduce Text2EQ, a human-in-the-loop semantic audio production interface that bridges the gap between intuitive language descriptors and equalization (EQ) parameters. Text2EQ enables users to describe their desired sound in natural language, mapping these descriptors to EQ settings. The system offers initial suggestions and supports iterative refinement, allowing users to adjust parameters through direct manipulation of the EQ parameters or additional natural language inputs. We aim to contribute to the current dialogue on the role and incorporation of intelligent audio tools into pre-existing workflows, highlighting the importance of balancing usability with creativity.

1. INTRODUCTION

Audio effects (FX) are indispensable for modern sound design and audio production. These FX are primarily utilized within digital audio workstations (DAWs).The control of these effects frequently involves intricate parameter adjustments on complex interfaces, demanding a high level of expertise to manipulate effectively. This complexity presents a significant learning curve, particularly for novice users and casual creators [1, 2]. Humans naturally describe sound using expressive, often subjective terms, such as "bright" or "crunchy." These descriptors, while accessible to non-experts, are not easily translated into the technical language of DSP-based interfaces like parametric equalizers. Recent advancements in multimodal deep learning models allow placing text descriptions into the same embedding space as audio examples (e.g. CLAP [3]). This offers a promising avenue for simplifying in audio production workflows by leveraging natural language as an intuitive interface for exploring the complex parameter spaces of audio FX.

Building off the idea of a shared audio-text embedding, a key research question emerges: can we build a human-centered interface that reduces the complexity of interacting with DSP-based audio effects by mapping text descriptions of effects to the parameters of the audio effects (e.g. reverb)? This interface could enable users to achieve their desired sound with greater ease by letting the user describe their goal in natural language. The interface would then suggest a set of effect parameters that the user can then tweak, either by using the traditional interface or by addition additional description. This goes far beyond traditional effect presets by carefully crafting a set of effect parameters for a *specific* piece of input audio, instead of a generic "one-size-fits-all" preset. Music production is inherently an iterative process [4, 5]. The proposed interaction is a human-in-the-loop interaction that can provide high-level guidance while letting users "get their hands dirty". Such interactions are well-suited to the audio production workflow and align ML-powered tools with users' artistic goals by engaging in a collaborative process.

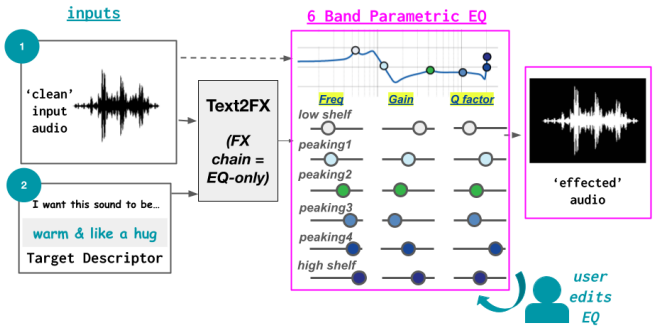

Figure 1. System Overview

In this late breaking demo, we introduce Text2EQ, a human-in-the-loop semantic audio production interface for EQ. With Text2EQ, users can describe their desired audio effect using natural language prompts such as "underwater" or "make brighter and powerful." The system then optimizes and maps these descriptors to suggested parameter settings, providing an initial ballpark approximation toward their desired sound. Crucially, Text2EQ allows for iterative refinement, enabling users to tweak parameters either directly or through additional natural language inputs to their liking, thus combining the power of both manual control of the traditional interface and ML-driven suggestions based on the user's description of desired goals.

2. MODEL OVERVIEW

We build on our previous work, Text2FX [6], an ML-based methodology where a target prompt (e.g., "bright") guides the selection of FX parameters. Randomly initialized settings are applied to an FX chain (e.g., EQ, EQ → Reverb,

Late-Breaking / Demo Session Extended Abstract, ISMIR 2024 Conference

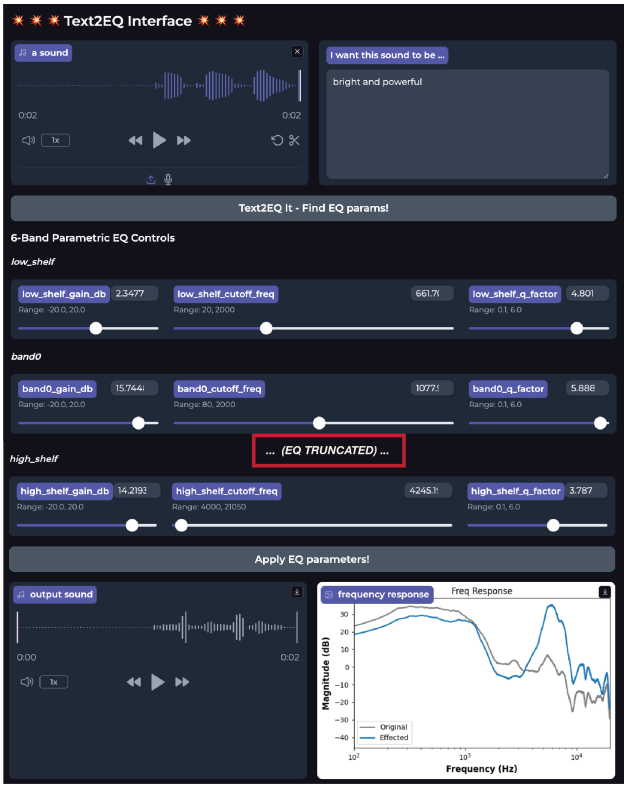

Figure 2. Screenshot of Text2EQ Interface

etc), and the processed audio, along with the prompt, is embedded into the CLAP [3] shared text/audio space. In Text2EQ we use the gradient-based optimization applied in Text2FX (minimizing cosine distance between text and audio embeddings), and iteratively adjust the FX parameters of a 6-band parametric EQ (implemented in dasp [1]) to align the audio embedding with the target text. The original Text2FX work only addressed the back-end machine learning problem and did not implement an interface or explore iterative human-in-the-loop interaction. The current work adds these elements to make an interactive co-creation interface.

3. INTERFACE DESIGN

We followed established design guidelines from prior research [5, 7, 8] for 1) interface design and 2) model selection & integration. **Simplicity:** Provide clear, intuitive controls for EQ parameters with a logical grouping and layout of elements to minimize information overload. **Flexibility and Customization:** Empower users by allowing them to modify or disregard outputs as needed, maintaining their sense of control and ownership. **Real-time (Multimodal) Feedback:** Offer both auditory (output) and visual feedback (frequency response) to facilitate user interaction and decision-making. **Balanced Unpredictability (Model):** Ensure a balance between the model's stochastic behavior and user expectations by selecting a configuration that minimizes output variation post-convergence, aligning with user needs.

[1] https://github.com/csteinmetz1/dasp-pytorch

Figure 1 shows the human-in-the-loop interaction paradigm and Figure 2 shows the Text2EQ interface prototype. Users can upload and play back their audio file and input a target prompt of any length via a free-text box. Pressing the button labeled "*Text2EQ It - Find EQ Params*" initiates the Text2FX optimization, returns the optimized EQ parameters after 600 iterations, and sets the EQ sliders set to the optimized parameters. To hear the effect of the optimized parameters, users press the *Apply EQ Parameters!* button, which generates (1) the modified audio with the EQ applied and (2) a visual comparison of the frequency responses between the original and modified audio. In the example (Figure 2), the highs around the 16kHz range were boosted to match the 'bright and powerful' target. Users can tweak the EQ sliders as needed, hear and see the results by pressing *Apply EQ Parameters!* repeatedly, allowing for continuous adjustment. The EQ sliders are additionally labeled with their respective ranges. The previews from the initial Text2EQ step remain unchanged, serving as grounding artifacts during iterations.

4. DISCUSSION

In our exploration of the preliminary Text2EQ interface, we identified several key areas warranting future development and enhancement. It is important to note that mapping a text description to EQ parameters is a one-to-many task: the same target effect (e.g. "make it brighter") could be accomplished in multiple ways (boosting high frequencies, cutting low frequencies, or doing both). The ML system should be designed such that the distribution of FX parameter predictions align with those chosen by human experts. Additionally, we emphasize the need for **low latency**. Though single-instance optimization eliminates the need for new training of a model when adding new or more FX, it takes approximately 40s to execute, delaying the provision of *immediate* feedback which adversely affects user experience.

One could also envision multiple Text2-*Effects* (e.g., Text2-*Reverb→Distortion*, Text2-*EQ→Reverb→Distortion*), which introduces the challenge of **FX chain generalization** and may involve designing a modular architecture that allows users to expand their FX setups and easily toggle FX on or off within an FX chain. This also raises interesting questions on the optimal integration of intelligent audio production tools. Should these tools be incorporated as plug-in assistants within existing software, or would they serve users better as standalone applications? Integrating into existing workflows could streamline operations for novices, making the technology more approachable while fostering a sense of familiarity. Conversely, a standalone tool may offer more specialized features that enhance the creative process for experienced professionals, potentially acting as a co-creation partner. Would this dual functionality successfully ease the learning curve of novices while retaining value as a co-creation tool for professionals? Ultimately, the integration method should focus on maximizing usability and enhancing the creative possibilities across all skill levels.

5. REFERENCES

[1] P. Seetharaman and B. Pardo, "Audealize: Crowdsourced audio production tools," *Journal of the Audio Engineering Society*, vol. 64, no. 9, pp. 683–695, 2016.

[2] A. T. Sabin, Z. Rafii, and B. Pardo, "Weighted-function-based rapid mapping of descriptors to audio processing parameters," *Journal of the Audio Engineering Society*, vol. 59, no. 6, pp. 419–430, 2011.

[3] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap: Learning audio concepts from natural language supervision," 2022. [Online]. Available: https://arxiv.org/abs/2206.04769

[4] A. Huang, H. V. Koops, E. Newton-Rex, M. Dinculescu, and C. J. Cai, "Ai song contest: Human-ai co-creation in songwriting," in *Proceedings of the 20th International Society for Music Information Retrieval Conference (ISMIR)*, 2020. [Online]. Available: https://program.ismir2020.net/poster_5-11.html

[5] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, "Novice-ai music co-creation via ai-steering tools for deep generative models," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–13.

[6] A. Chu, P. O'Reilly, J. Barnett, and B. Pardo, "Text2fx: Harnessing clap embeddings for text-guided audio effects," 2024. [Online]. Available: https://arxiv.org/abs/2409.18847

[7] A. Roberts, J. Engel, Y. Mann, J. Gillick, C. Kayacik, S. Nørly, M. Dinculescu, C. Radebaugh, C. Hawthorne, and D. Eck, "Magenta studio: Augmenting creativity with deep learning in ableton live," 2019.

[8] C. Kayacik, S. Chen, S. Noerly, J. Holbrook, A. Roberts, and D. Eck, "Identifying the intersections: User experience+ research scientist collaboration in a generative machine learning interface," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–8.